Reviewer 1

Tabel 2 er umiddelbart ikke klart gennemskuelig. Kan måske stilles op på en sådan måde, at den bliver lettere tilgængelige/ forståelige (tabel 1 er let forståelig).

God velskreven artikel, men jeg mener der bør arbejdes mere med specielt tabel 2, som umiddelbart ikke er let tilgængelig for læsere med en ikke så stor forståelse for statistik/ læsere der ikke er vant til at anvende statistik.

LON: Tak for de opmuntrende ord. Vi tror desværre ikke, det er hverken muligt eller tilrådeligt. Det er simpelthen standarden/normen, når man rapporterer regressionsmodeller, som vi har fulgt. Det vi har gjort, som et forsøg på at imødekomme læsevenligheden, er at lægge to søjler (OR og 95% CI) sammen, så der er færre søjler at forholde sig til. Men hvis I har et konkret alternativt og bedre forslag, lytter vi selvfølgelig rigtig gerne til det.

Savner en figur af Tinto's dropout model, der bliver refereret en del til. Er det muligt at den kan indgå i artiklen.

LON: God pointe, det savnede vi faktisk også. Vi har tilføjet en modificeret udgave af Tinto’s model, hvor optagelses tests (’program specific admission testing’) er tilføjet. Modificeret, fordi hvis vi skulle have reproduceret Tinto’s helt originale model ville Sagepub have os til at betale 34,04 EUR for hver 6. måned figuren er frit tilgængelig på internettet, som den jo er i DUT. Det anså vi ikke for holdbart.

Savner desuden henvisning til litteratur der omhandler kvinder og idræt (før underoverskriften Test-based admission i afsnittet discussion).

LON: Det kan vi godt forstå. Det er hermed rettet.

Savner en diskussion af statistisk behandling af en så lille gruppe som 47 studerende der fik adgang pga karakterter. Er gruppestørrelsen passende at lave statistik på, eller kan størrelsesordnene have indflydelse på de følgende målte parametre (drop-out).

LON: Det som er afgørende for, om man kan lave multivariate logistiske regressionsmodeller er antallet af frafaldne (dropouts) i den samlede population. Tommelfingerreglen er, at man helst ikke skal inkludere mere end 1 variabel pr. 10-20 frafaldne i den multivariate model – så kan man risikere at skræve videre end bukserne kan holde. Vi har 79 frafaldne i alt og 4 variable i den multivariable model, så det skulle være i orden. Desuden har vi haft statistiker Lars Korsholm (dengang SDU) til at tjekke den endelige model, og han har sagt god for den. Du har ret i at kvote 1 gruppen er forholdsvis lille i forhold til kvote 2 gruppen (47 vs. 402). Men det er faktisk sådan, at forskellen på fordelingen af frafaldne i de to grupper er statistisk signifikant. Det betyder at vi kan anse frafaldsprocenterne i de to grupper for ikke at være en statistisk tilfældighed, selv om den ene gruppe er mindre. Prøv selv at regne på p-værdien for 2 x 2 tabellen her (data er fra resultatafsnittet):

|  |  |  |  |
| --- | --- | --- | --- |
|  | Frafaldne | Ikke-frafaldne | I alt |
| Kvote 1 | 14 | 33 | 47 |
| Kvote 2 | 65 | 337 | 402 |
| I alt | 79 | 370 | 449 |

Jeg får p=0,021 med Fisher’s Exact Test - får du ikke noget lignende? Man kan også regne OR for at falde fra som kvote 1’ er ud via tabellen, den er: (14 x 337)/(33 x 65)=2,20. De tilsvarende tal (OR og p) genfinder du i tabel 2 (’Grade-based admission’, univariate analyses), bare udregnet med logistisk regression. Med andre ord: Vi ville simpelthen ikke kunne finde signifikante OR, hvis vores sample af kvote 1’ere havde været for lille.

Hvis vi derimod *ikke* havde fundet en effekt af optagelsesprøver (som vi jo har) kunne du med rette spørge, om vi havde lavet en type 2 fejl og betvivlet vores power. Men det er jo lige tværtom her. Vi fandt en forskel, selv med så lille et sample af kvote 1’ere. Det er faktisk meget stærkt. Og effekten er ovenikøbet ganske stabil, selv når vi kontrollerer for andre indflydelsesrige prædiktorvariable i den multivariate model (tjek selv tabel 2 for de negligerbare ændringer i OR og p-værdi for ’grade-based admisssion’ for hhv. univariate vs. multivariate analyser).

Vi mener som udgangspunkt ikke, at vi skal bruge dyrebar artikelplads til trivielle statistiske betragtninger, specielt ikke, fordi de ikke giver anledning til reelle statistiske problemer i studiet her. Det håber vi at I kan finde forståelse for. Vi tror ikke det ville gavne læsevenligheden heller: vi kan simpelthen ikke give et statistik-kursus og yde studiet retfærdighed i samme artikel, med den ordgrænse som pt. eksisterer. Derfor vil vi simpelthen foreslå, at vi *ikke* diskuterer den statistiske behandling af de 47 kvote 1’ere yderligere i artiklen. Derved undgår vi både at kede dem, der kan gennemskue statistikken og forvirre dem som ikke kan.

Det ville være interessant at sammenligne med frafaldsprocent fra tidligere årgange som et supplement til nærværende undersøgelse af 402 vs 47 studerende (det er jo praktisk talt kun 8 stud pr år der kommer igennem pga snit pr årgang fra 2002 til 2007).

LON: Som nævnt herover - vi fandt en effekt af optagelsesprøver, *selv* med et mindre sample af kvote 1’ere, en effekt som endda overlevede kontrol for confounding. Det har ikke været muligt for os, at opdatere populationen, vi har skam forsøgt bare for en god ordens skyld. Det strandede dels på problemer med administrative ressourcer til at levere dataudtrækket, samt at studiekontoret ikke længere kunne levere en sammenlignelig udtrækningsprocedure. Det ville være endnu bedre om andre forskere vil undersøge andre populationer på andre institutioner i fremtiden for den samlede viden om optagelsesprøver og frafald.

Reviewer 2

Major - very relevant and well-written paper. some minor/specific comment below.

LON: Thank you.

Specific comments

Methods: "The other ca. 80% of available.." - "ca" should be "approximately"

LON: We agree, and have made the suggested change.

It is not clear why a "General knowledge test" was chosen rather than a test of knowledge within the relevant field (sport sciences) - this could/should be explained in more details.

LON: We explained it in the method section with the words: “The aim of this [General Knowledge] test was to secure broad societal and cultural interests and curiosity towards knowledge in general amongst the ‘best match’ students, i.e. the specific content of the test was not thought to be particularly relevant.”

 We do not have more details available on this choice than what is already described here.

"generalizability" ?!?

LON: ‘Generalizability’ coefficients are extensions of classical reliability coefficients, which allows for the disentanglement of more than one source of error within the same single coefficient. This extension of Classical Test Theory is called Generalizability Theory, and the term ‘generalizability’ was therefore used in brackets after the word ‘reliability’. We have removed the word ‘generalizability’ now to ease the general understanding, because we think your confusion is understandable and probably shared by other readers.

RESULTS - p=0.000 ?!? - suggest you use standard categories - here P < 0.001

LON: In our experience, p-value estimates are typically reported as exact values with 3 decimals in the international literature. Modern statistical packages like STATA supply *exact* value estimates for p. This means that p=0.000 is in fact an exact value below 0.0005. We believe that ‘p<0.001’ is a leftover from the old days, when we had to use p-value intervals (i.e. p<0.05, p<0.01 or p<0.001), because we had to consult distributional tables in textbooks to get an estimate of p. In any case, p<0.001 is actually less precise than p=0.000, so we suggest that we keep it precisely as it is.

However, your comment drew our attention to the fact, that we have used , instead of . in the values in tables, so we have corrected this mistake.

"In other words: the ‘best match’ students were less well-qualified academically than the traditionally admitted students (grade-based admission)." - this is disussion and should be moved to that section.

LON: OK, we have removed the sentence completely now. It was meant to clarify the results described in the previous line to readers with little statistical understanding.

Similarly for the last part of this sentence - "The odds of dropping out were 2.22 times larger for students in the grade-based admission group compared to students in the best-match admission group though the latter were less well-qualified academically "

LON: All right then – we have removed the last part of the sentence, though we tend to disagree somewhat. It is not for discussion per se, it is a result.

Discussion:

"We found that pre-university grades were not predictors of subsequent dropout for these particular students in this particular setting, which is in contrast to global findings in higher education" - could this relate to the fact that you only observe student with a grading above a certain cut-off. I.e. the students you obtain have a sufficient level ?

LON: Yes and no. ‘Yes’ because: other Danish authors (Christensen & Juul, 1999) have found that having obtained medium average grades in upper secondary school (8 or above on the old grading scale) was protective of dropout in medical school at AU. Most likely ‘no’ in this particular case, because: our sample of 449 sports science students had a relatively wide range of grades (see ‘pu-GPA’ in table 1), with a minimum of 2.6 and a maximum of 10.8, i.e. we do not just observe students with medium-high grades (‘above a certain cut-off’) – we actually have students with quite low grades in our sample too (2.6 is just above the pass mark 2). This means that the risk of ‘restriction of range’ is even less likely to minimized the chances of finding a significant association, compared to what it often the case when examining more academically ‘elite’ students (e.g. medical students), and that we may therefore put some serious trust in the result (no effect of pu-GPA on dropout in year 1 & 2 sports science) for this sample. It is perhaps more likely, that year 1 and 2 of this sports science curriculum is just not that demanding academically – but perhaps more demanding of practical sports skills, which do constitute an important part of the admission test. We have therefore added the following thought to the discussion of grades:

“An explanation for the result that prior grades did not predict dropout while program specific admission testing did, could be if year 1 and 2 of the sports science curriculum were less demanding academically, but perhaps more demanding of practical sports skills, which also constitute an important part of the admission test.”

This brings the word count up to 24,675 characters incl. spaces (excluding the abstract).